Manual de Utilização do Etiquet(H)AREM

Paula Carvalho
Linguateca – Pólo de Lisboa no XLDB/LasiGE/FCUL

Hugo Gonçalo Oliveira
Linguateca - Pólo de Coimbra, CISUC

Versão 1.0 (29 de Abril de 2008)

O **Etiquet(H)AREM** é uma ferramenta de auxílio à anotação de *corpora*, concebida por Hugo Gonçalo Oliveira, para a etiquetagem de Entidades Mencionadas (EMs) e de relações entre EMs, no âmbito do HAREM (http://www.linguateca.pt/HAREM/).

Requisitos básicos na utilização do programa

(i) A utilização desta ferramenta pressupõe a instalação de uma máquina de **JAVA** - Java Runtime Environment (JRE) 1.6 ou mais recente (http://www.java.com/en/download/manual.jsp).

(ii) O ficheiro a ser anotado tem de estar em formato **xml**, caso contrário o programa não o abre.

(iii) Só são suportados ficheiros XML com DTDs, se estas forem externas. Nesse caso, o ficheiro **.dtd** terá de se encontrar na mesma directoria para onde o DOCTYPE estiver a apontar. No caso de o ficheiro ter uma DTD interna, não há garantias de bom funcionamento do programa.

(iv) Os valores possíveis para os atributos das EMs estão compreendidas no ficheiro **harem3.conf** (*cf.* Anexo).

Lista de notações a utilizar

O ficheiro *harem3.conf* corresponde à listagem das Categorias (**C**), e respectivos tipos (**T**) e/ou subtipos (**S**), previstos no âmbito das Directivas do Segundo Harem. O referido ficheiro pode incluir ainda outros atributos igualmente tidos em consideração na anotação (caso de (**X**) e (**Y**), como abaixo referido), bem como as relações (**R**) previstas entre as EMs.

Para adicionar uma nova categoria, tipo ou subtipo, basta introduzir o respectivo nome (em maiúsculas), antecedido de **C:** , **T:** ou **S:** , respectivamente. No que respeita à categoria TEMPO, é ainda possível adicionar os atributos **X:** (TEMPO_REF) e **Y:** (SENTIDO), ambos relativos ao subtipo DATA.

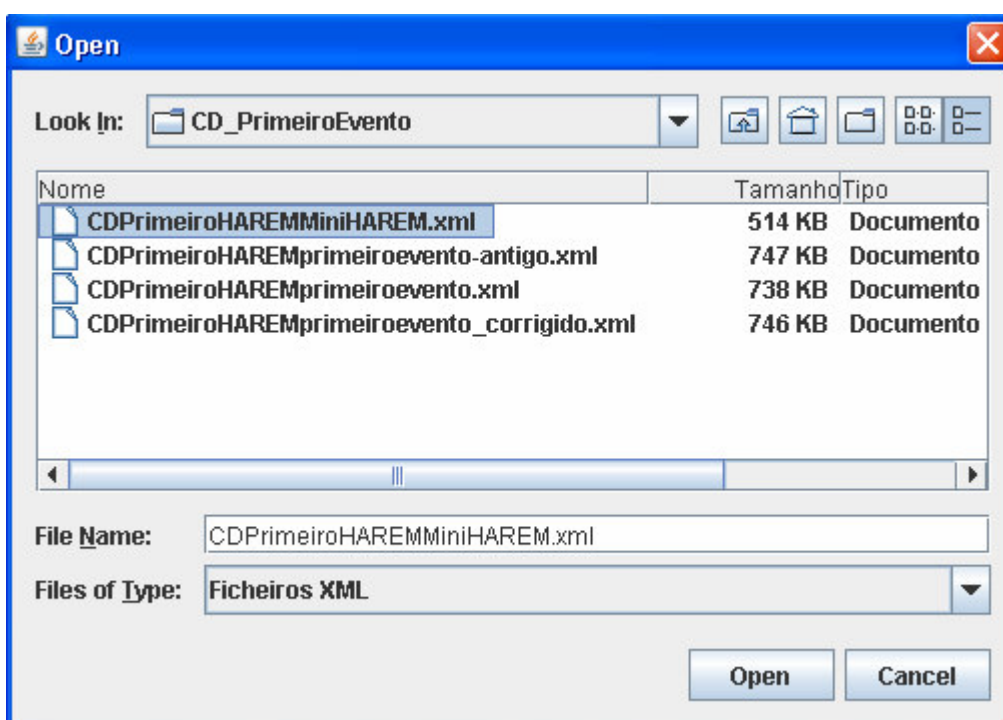
Os atributos *categoria*, *tipo* e *subtipo* (e eventuais '*subsubtipos*') encontram-se organizados hierarquicamente, por esta ordem. Assim, sempre que se insere uma entrada do tipo *T:xxx*, a categoria a que esse tipo pertence corresponderá à entrada

C:yyy mais próxima e imediatamente acima de *T:xxx*. Os subtipos funcionam de forma idêntica.

Para especificar os tipos de Relações (**R**) entre EMs, basta declará-las a seguir a **R:**. Neste caso, convencionou-se que as relações seriam grafadas em minúsculas, ao contrário das categorias, dos tipos e dos subtipos, que são grafados em maiúsculas.

Manuseamento do programa propriamente dito

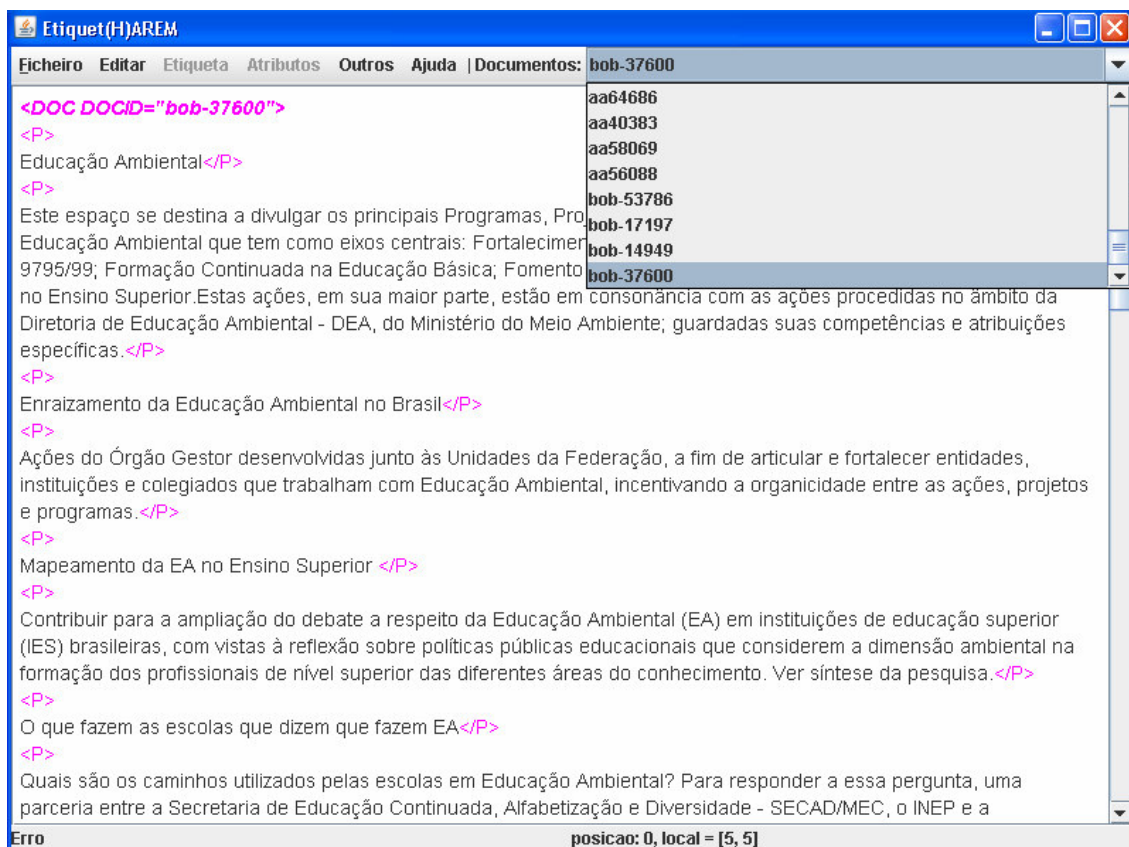
Iniciar o Etiquet(H)arem:

a) Clicar duas vezes sobre a aplicação etiquet(h)arem.jar, ou, em alternativa,
b) Abrir explicitamente o programa numa consola: java –jar etiquetharem.jar

Obs: Ao abrir a aplicação, é imediatamente pedido para seleccionar o ficheiro a anotar.

Menus do Etiquet(H)arem

(i) **Ficheiro**

ABRIR - Permite abrir um (novo) ficheiro.
GUARDAR - Permite gravar o ficheiro de trabalho.
GUARDAR COMO - Permite atribuir um novo nome ao ficheiro de trabalho.
TERMINAR – Permite sair da aplicação.

Obs: Sempre que um dado ficheiro é aberto, à frente de **Documentos** aparece uma lista que é preenchida com o DOCID de todos os documentos (DOC) do ficheiro. Incialmente é mostrado o primeiro documento, mas é possível visualizar qualquer documento dessa listagem, através da selecção do DOCID correspondente.

(ii) **Editar**

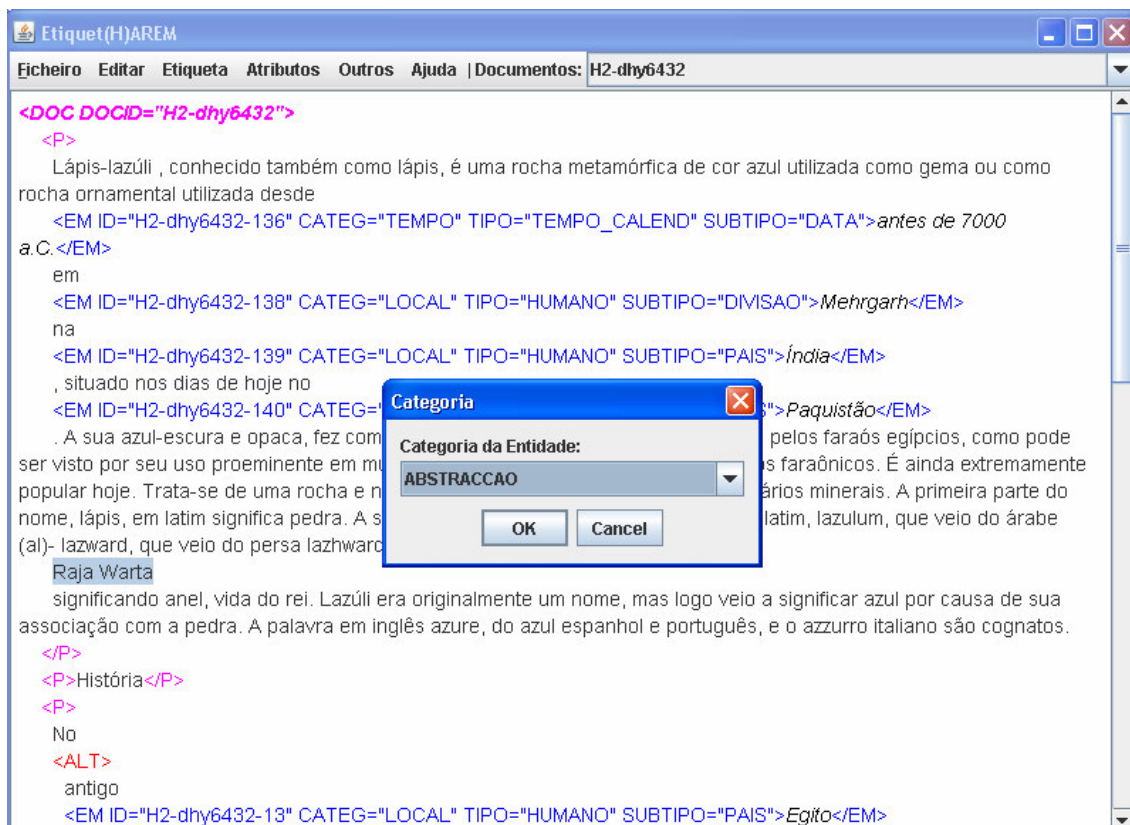
Os comandos compreendidos neste menu são idênticos aos utilizados na generalidade das aplicações.

ANULAR – permite anular uma operação. No entanto, uma operação para o programa pode não ser o mesmo do que uma operação para o utilizador; por exemplo, a anulação de uma etiqueta inserida implica a repetição do comando.

REPETIR - permite repetir a operação anulada pelo comando anterior.

CUT-TO-CLIPBOARD, **COPY-TO-CLIPBOARD** e **PASTE-FROM-CLIPBOARD**, comandos que permitem, respectivamente, cortar um fragmento do texto, copiar um fragmento do texto ou adicionar um fragmento ao texto.

(iii) **Etiqueta**

EM – Serve para atribuir uma etiqueta a uma palavra ou sequência de palavras previamente seleccionadas no texto.

EM VAGA – Deve ser utilizado para etiquetar EMs que possam ser vagas entre 2 ou mais categorias, tipos e/ou subtipos. A vagueza é representada através do caracter "|". Ao seleccionar esta funcionalidade, o programa pede para escolher o índice de vagueza, ie., o número de etiquetas (ou interpretações) diferentes que a referida EM poderá receber (2, 3, 4, 5, 6).

EM ALTERNATIVA – Esta funcionalidade permite atribuir duas ou mais análises alternativas a uma mesma sequência de palavras previamente seleccionadas no texto. Neste caso, o programa repetirá o fragmento do texto seleccionado tantas vezes quanto o número de análises alternativas seleccionadas (2, 3, 4, 5, 6). As diferentes análises encontram-se separadas através do caracter "|", e o fragmento do texto onde existem análises alternativas está delimitado, à esquerda e à direita, pelas etiquetas <ALT> e </ALT>, respectivamente.

REPETIR, REMOVER e ALTERAR – Estes comandos permitem, respectivamente, repetir, remover ou alterar uma etiqueta previamente atribuída a uma dada EM. Para isso, basta seleccionar toda a etiqueta e proceder às alterações desejadas.

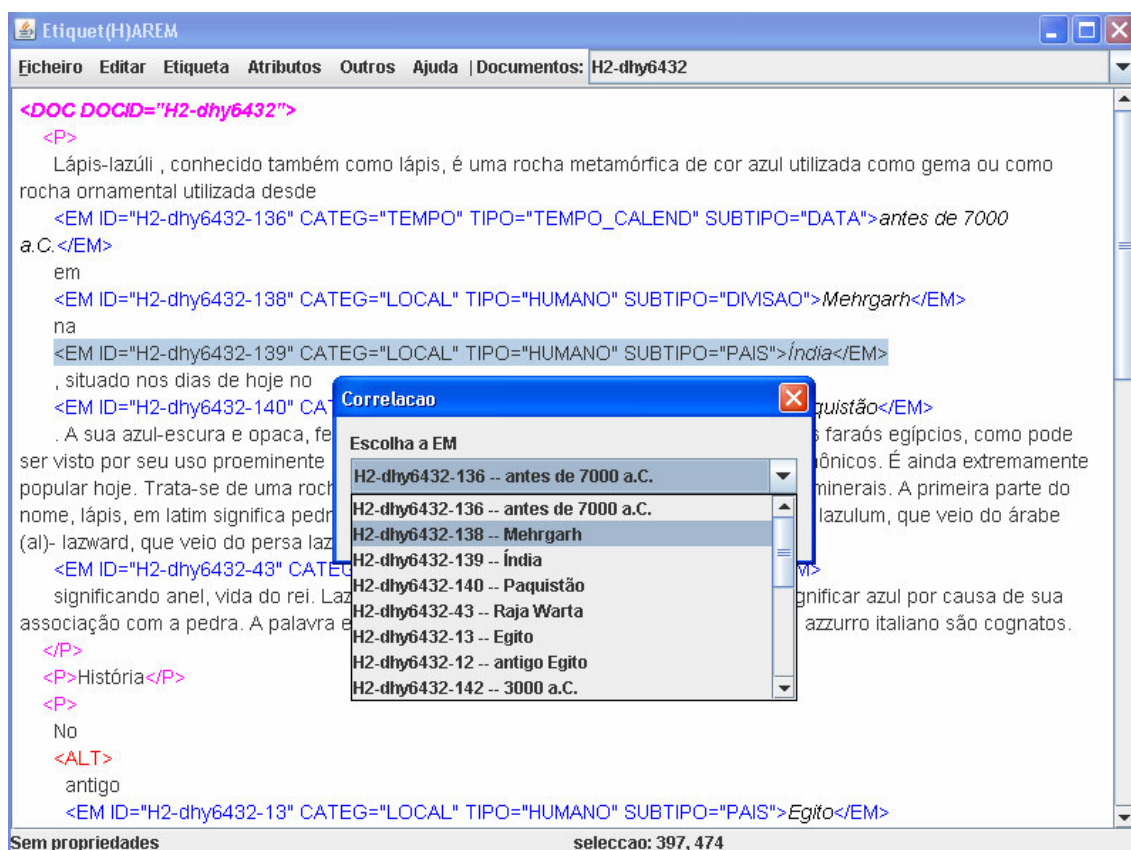
AUMENTAR VAGUEZA – Esta funcionalidade permite atribuir uma nova análise a uma EM previamente etiquetada no texto. Para isso, basta seleccionar toda a etiqueta associada a essa EM e introduzir os novos atributos desejados.

NOVA ALTERNATIVA – Esta funcionalidade permite introduzir uma nova análise alternativa a uma EM previamente etiquetada no texto com duas ou mais análises alternativas. Neste caso, o programa apenas reproduz um novo bloco de texto, sem qualquer notação, para posterior etiquetagem.

OMITIR – Esta funcionalidade permite marcar um fragmento de texto como “omitido”, colocando-o entre as etiquetas <OMITIDO> </OMITIDO>. O texto omitido não será alvo de avaliação.

(iii) **Atributos**

Este menu serve fundamentalmente para adicionar nova informação a uma dada EM que já tenha sido anteriormente etiquetada.

CORRELAÇÃO – Permite inserir o tipo de relação que uma dada EM mantém com uma outra EM. Para isso é necessário seleccionar antes a EM e respectiva etiqueta. Será depois mostrada uma lista com todas as EMs já anotadas dentro do mesmo documento, de forma a que o utilizador possa escolher aquela com que existe a relação. Depois disso, será pedido que se seleccione o tipo de relação.

TIPO e **SUBTIPO** – Estas funcionalidades permitem inserir o TIPO e/ou SUBTIPO a uma EM do texto cuja etiqueta não contenha esses atributos. Para isso, é necessário seleccionar antes a EM e respectiva etiqueta. Será depois mostrada uma lista com as possibilidades que estes campos podem ter.

TEMPO – Permite inserir os atributos TEMPO_REF e SENTIDO, relativos à categoria TEMPO tipo TEMPO_CALEND subtipo DATA.

COMENTÁRIO – Permite inserir o atributo *comentário* (COMENT) na EM. É necessário ter algum cuidado na sua utilização já que este atributo se pode inserir em qualquer parte do texto, sendo, no entanto, válido apenas quando se encontra dentro de uma etiqueta de EM. O atributo *comentário* pode ser utilizado pelo anotador para acrescentar algo à sua anotação, por exemplo, a indicação de que não tem a certeza se a mesma foi bem feita.

META ERRO – Trata-se de uma funcionalidade que, por enquanto, não está a ser usada. Foi implementada sobretudo para dar conta de (*cf.* http://poloxldb.linguateca.pt/harem.php?l=classificacao_v3_sem):

«casos em que há enganos de ortografia ou grafia no texto, em particular quando uma palvara tem uma maiúscula a mais ou a menos e tal é notório, escolhemos corrigir mentalmente a grafia (maiúscula /minúscula) de forma a poder classificar correctamente. Além disso, estamos a pensar em marcar estes casos, na colecção dourada, com uma classificação `META="ERRO"`.

Certo : O grupo terrorista <PESSOA TIPO="GRUPO" META="ERRO">Setembro negro</PESSOA>»

(iii) **Outros**

Este menu compreende uma série de comandos que envolvem a manipulação e apresentação do próprio texto:
LOCALIZAR – Permite identificar uma palavra ou sequência de palavras no texto do documento que se está a visualizar.

MOSTRA ETIQUETAS e **ESCONDE ETIQUETAS** – permitem a visualização do texto com ou sem etiquetas, respectivamente.

VALIDAR XML – permite fazer uma validação do XML, tendo em conta (se existir) a DTD.

TAMANHO DA LETRA – permite aumentar ou diminuir o tamanho da letra do texto visualizado.

(iv) **Ajuda**

COMO ETIQUETAR – Explica os diferentes modos de atribuição de uma etiqueta ou atributo a uma dada EM no texto.

ACERCA – Dá a indicação do programa e respectiva versão que se está a utilizar.

Agradecimentos
Queremos agradecer à Diana Santos e à Cláudia Freitas as importantes sugestões a versões preliminares deste documento. Este trabalho foi desenvolvido no âmbito do projecto Linguateca contrato nº 339/1.3/C/NAC, financiado pelo governo português e pela União Europeia.

Anexo: **harem3.conf**

#Categorias (**C**), *Tipos* (**T**) *e Subtipos* (**S**)
#TEMPO_REF (**X**), *SENTIDO* (**Y**)
#Tipos de referencia (**R**)

C:PESSOA
T:INDIVIDUAL
T:CARGO
T:GRUPOCARGO
T:GRUPOMEMBRO
T:MEMBRO
T:GRUPOIND
T:POVO
T:OUTRO
C:ORGANIZACAO
T:ADMINISTRACAO
T:EMPRESA
T:INSTITUICAO
T:OUTRO
C:LOCAL
T:HUMANO
S:PAIS
S:DIVISAO
S:REGIAO
S:CONSTRUCAO
S:RUA
S:OUTRO
T:FISICO
S:AGUACURSO
S:AGUAMASSA
S:RELEVO
S:PLANETA
S:ILHA
S:REGIAO
S:OUTRO
T:VIRTUAL
S:COMSOCIAL
S:SITIO
S:OBRA
S:OUTRO
C:OBRA
T:REPRODUZIDA
T:ARTE
T:PLANO
T:OUTRO
C:ACONTECIMENTO
T:EFEMERIDE
T:ORGANIZADO
T:EVENTO
T:OUTRO
C:ABSTRACCAO
T:DISCIPLINA
T:ESTADO
T:IDEIA
T:NOME
T:OUTRO
C:COISA
T:CLASSE
T:SUBSTANCIA
T:OBJECTO
T:MEMBROCLASSE
T:OUTRO
C:VALOR
T:CLASSIFICACAO
T:QUANTIDADE
T:MOEDA
T:OUTRO
C:OUTRO
T:OUTRO
C:TEMPO
T:CALENDARIO
S:DATA
X:ABSOLUTO
X:TEXTUAL
X:ENUNCIACAO
Y:ANTERIOR
Y:POSTERIOR
Y:ANTERIOR_OU_SIMULT
Y:POSTERIOR_OU_SIMULT
S:INTERVALO
S:HORA
T:DURACAO
T:FREQUENCIA
T:GENERICO
#Tipos de referencia(R)
R:ident
R:incluido
R:inclui
R:ocorre_em
R:sede_de
R:outro